IN THE CLAIMS:

1. (Previously Presented) A program storage device readable by a machine, tangibly embodying a program of instructions executable by the machine to perform method steps for speech synthesis, the method steps comprising:

providing a text string comprising a plurality of words and phonemes and corresponding spoken audio signal;

extracting acoustic feature data from said audio signal;

aligning the text string and the acoustic feature data and outputting a set of duration contours indicative of the duration of each word and phoneme;

extracting pitch contour parameters from said audio spoken input;

automatically generating a marked-up text corresponding to the spoken utterance using the pitch and duration contours; and

generating a synthetic waveform using the marked-up text.

2-3. (Canceled)

4. (Previously Presented) The program storage device of claim 1, wherein the instructions for aligning comprise instructions for segmenting said spoken audio signal into time-segmented regions, wherein each time-segmented region is mapped to a corresponding phoneme.

5. (Previously Presented) The program storage device of claim 1, wherein the alignment is performed using a Viterbi alignment process.

6. (Canceled).

7. (Previously Presented) The program storage device of claim 1, wherein the instructions for automatically generating a marked-up text comprise instruction for directly specifying the pitch contour and duration parameters as attribute values for mark-up elements.

8. (Previously Presented) The program storage device of claim 1, wherein the instructions for automatically generating a marked-up text comprise instructions for assigning abstract labels to the pitch contour and duration parameters to generate a high-level markup.

9. (Original) The program storage device of claim 1, wherein the marked-up text is generated using SSML (speech synthesis markup language).

10. (Original) The program storage device of claim 1, further comprising instruction for processing phonetic content of the spoken utterance to generate the synthetic waveform having a desired pronunciation.

11. (Previously Presented) A method for speech synthesis, comprising the steps of:

providing a text string comprising a plurality of words and phonemes and corresponding spoken audio signal;

extracting acoustic feature data from said audio signal;

aligning the text string and the acoustic feature data and outputting a set of duration contours indicative of the duration of each word and phoneme;

extracting pitch contour parameters from said audio spoken input;

automatically generating a marked-up text corresponding to the spoken utterance using the pitch contour and duration parameters; and

generating a synthetic waveform using the marked-up text.

12-13. (Canceled)

14. (Previously Presented) The method of claim 11, wherein aligning comprises extracting acoustic feature data from the spoken utterance and time-aligning the spoken input to the corresponding text string using the acoustic feature data.

15. (Previously Presented) The method of claim 11, wherein aligning is performed using a Viterbi alignment process.

16. (Canceled)

17. (Previously Presented) The method of claim 11, wherein automatically generating a marked-up text comprises directly specifying the pitch contour and duration parameters as attribute values for mark-up elements.

18. (Previously Presented) The method of claim 11, wherein automatically generating a marked-up text comprises assigning abstract labels to the pitch contour and duration parameters to generate a high-level markup.

19. (Original) The method of claim 11, wherein the marked-up text is generated using SSML (speech synthesis markup language).

20. (Original) The method of claim 11, further comprising processing phonetic content of the spoken utterance to generate the synthetic waveform having a desired pronunciation.

21. (Previously Presented) A text-to-speech (TTS) system, comprising:

a prosody analyzer for determining prosodic parameters of a spoken utterance corresponding to an input text string and automatically generating a marked-up text

corresponding to the spoken utterance using the prosodic parameters, wherein the prosody analyzer comprises:

an acoustic feature extraction module that extracts acoustic feature data from said spoken utterance;

an alignment module for aligning the input text string with the spoken utterance using said acoustic feature data to generate duration contour information of elements comprising the input text string;

a pitch contour extraction module for determining pitch contour information for the spoken utterance; and

a conversion module for including markup in the input text string in accordance with the duration and pitch contour information to generate the marked up text; and

a TTS system for generating a synthetic waveform using the marked-up text.

22. (Original) The system of claim 21, further comprising a user interface that enables a user to input the spoken utterance and input a text string corresponding to the spoken utterance.

23. (Original) The system of claim 21, wherein the prosody analyzer processes phonetic content of the spoken utterance to generate the synthetic waveform having a desired pronunciation.

24. (Canceled)

25. (Previously Presented) The program storage device of claim 1, wherein extracting acoustic feature data from said audio signal comprises digitizing the spoken audio signal into a set of frames and transforming the digitized input waveforms into a set of feature vectors on a frame-by-frame basis.

26. (Previously Presented) The program storage device of claim 25, wherein transforming the digitized input includes producing a 24-dimensional cepstra feature vector for every 10ms of the spoken audio signal, concatenating frames to the left and to the right of a current frame to augment a current cepstral vector, and reducing each augmented cepstral vector to a 60-dimensional feature vector using linear discriminant analysis.

27. (Previously Presented) The method of claim 11, wherein extracting acoustic feature data from said audio signal comprises digitizing the spoken audio signal into a set of frames and transforming the digitized input waveforms into a set of feature vectors on a frame-by-frame basis.

28. (Previously Presented) The method of claim 27, wherein transforming the digitized input includes producing a 24-dimensional cepstra feature vector for every 10ms of the spoken audio signal, concatenating frames to the left and to the right of a current frame to augment a current cepstral vector, and reducing each augmented cepstral vector to a 60-dimensional feature vector using linear discriminant analysis.